Genetic Epidemiology 18:293–306 (1999)

Score Tests for Familial Correlation in Genotyped-Proband Designs

Raymond J. Carroll,[1]* Mitchell H. Gail,[2] Jacques Benichou,[3] and David Pee[4]

[1]*Department of Statistics, Texas A&M University, College Station, Texas*
[2]*Biostatistics Branch, Division of Cancer Epidemiology and Genetics, National Cancer Institute, Bethesda, Maryland*
[3]*Biostatistics Unit, University of Rouen Medical School, CHU de Rouen, France*
[4]*Information Management Systems, Rockville, Maryland*

In the genotyped-proband design, a proband is selected based on an observed phenotype, the genotype of the proband is observed, and then the phenotypes of all first-degree relatives are obtained. The genotypes of these first-degree relatives are not observed. Gail et al. [(1999) Genet Epidemiol] discuss likelihood analysis of this design under the assumption that the phenotypes are conditionally independent of one another given the observed and unobserved genotypes. Li and Thompson [(1997) Biometrics 53:282–293] give an example where this assumption is suspect, thus suggesting that it is important to develop tests for conditional independence. In this paper, we develop a score test for the conditional independence assumption in models that might include covariates or observation of genotypes for some of the first degree relatives. The problem can be cast more generally as one of score testing in the presence of missing covariates. A standard analysis would require specifying a distribution for the covariates, which is not convenient and could lead to a lack of model-robustness. We show that by considering a natural conditional likelihood, and basing the score test on it, a simple analysis results. The methods are applied to a study of the penetrance for breast cancer of BRCA1 and BRCA2 mutations among Ashkenazi Jews. Genet. Epidemiol. 18:293–306, 2000. © 2000 Wiley-Liss, Inc.

Key words: asymptotics; epidemiology; genetics; generalized linear mixed models; kin-cohort design; proband; random effects; retrospective studies; score tests

Contract grant sponsor: National Cancer Institute; Contract grant number: CA-57030; Contract grant sponsor: National Institute of Environmental Health Sciences; Contract grant number: P30-ESO9106.

*Correspondence to: R.J. Carroll, Department of Statistics, Texas A&M University, College Station, TX 77843-3143. E-mail: carroll@stat.tamu.edu

Received 11 September 1998; Accepted 15 June 1999

INTRODUCTION

Once a gene has been identified as affecting risk of disease and the gene has been isolated, there are several study designs available to determine the risk of the disease (the penetrance) from that gene in the general population, including cohort and case-control designs. A promising design is the genotyped-proband design, which was used by Struewing et al. [1997] to estimate the cumulative risk of developing breast cancer in Ashkenazi Jewish women who carry mutations of BRCA1 and BRCA2. Wacholder et al. [1998] discuss this design and called it the "kin-cohort" design, and Gail et al. [1999] constructed likelihoods and described calculations of sample sizes needed to achieve required precision of penetrance estimates.

In this design, a proband is selected and a phenotype obtained, perhaps at random from the population of cases with disease, or at random from the population of non-cases. For quantitative traits such as blood pressure, "cases" might have a prescribed elevation. After observing the phenotype of the proband, the proband's genotype is obtained, as well as the phenotypes of all first-degree relatives. Occasionally, one might also obtain the genotype of a randomly selected first-degree relative. A major factor is that the genotypes of most of the family are unobserved, and hence missing.

Under assumptions of simple Mendelian genetics with two alleles (wild-type a or mutant A), Gail et al. [1999] construct likelihoods for parameters in models relating phenotypes to genotypes, and for allele frequency. Their analysis makes a crucial assumption, namely that within a family, the *H*s are independent of one another given only the genotypes. Violations of this assumption would affect the validity of both maximum likelihood estimates and likelihood inferences. Recently, in analyzing a different design where no genotypes are observed, Li and Thompson [1997] describe an example in a survival analysis context where full random effects analysis suggests that the conditional independence assumption fails. It is thus of considerable interest to derive methods to test this assumption.

Indeed, it is likely that there are residual family effects that could induce familial correlations in breast cancer risk conditioned on the status of major cancer genes, such as BRCA1 or BRCA2. As Li and Thompson state, "Often, dependence among relatives is also due to random effects such as shared environment within a family." Diet, endogenous hormones, and reproductive history are thought to influence breast cancer risk, and these factors are not known to be affected by BRCA1 or BRCA2. Because these factors may be correlated within families as a result of other genetic influences or learned behaviors, it is plausible that the assumption of conditional independence might fail.

This paper proposes a score test of the conditional independence assumption, in a context that generalizes previous work to allow for family-level and individual-level covariates. The test, a score test for a random effect, exploits the work of Liang [1987], but differs from that work because we incorporate the missing genotypes. For other recent important work on score tests, see Commenges et al. [1994, 1995] and Commenges and Jacqmin-Gadda [1997]. A recent review of these tests is given by Lin [1997].

Score tests are often used to assess the hypothesis of independence of observations within a family, or more generally within a cluster. Score tests have the advantages of computational simplicity and model robustness. For example, suppose that

all genotypes were observed, that the gene is autosomal dominant, and that the H is binary. Let $g = 1,2,3$ correspond, respectively, to genotypes AA, Aa, and aa. With no covariates, Gail et al. [1998] assume no random effect and write $\text{pr}\{\text{disease}|g \in (1,2)\} = H(\gamma)$ and $\text{pr}(\text{disease}|g = 3) = H(\eta)$, where $H(v) = 1/\{1 + \exp(-v)\}$ is the logistic distribution function. The value of $H(\gamma)$ is the "penetrance" of the dominant mutant allele, and $H(\eta)$ is the penetrance of the wild type. To test for family-level correlation, one might first hypothesize a random effect ζ with mean zero and variance κ, and that within a family, $\text{pr}\{\text{disease}|g \in (1,2)\} = H(\gamma + \zeta)$ and $\text{pr}(\text{disease}|g = 3) = H(\eta + \zeta)$. A score test is a test of $\kappa = 0$, with the parameters (γ, η) estimated under the null hypothesis. Score tests have the advantage that one need not specify a *distribution* for the random effect ζ, and they are locally most powerful tests. The distributional-robustness of a score test is a powerful feature, as is the typical ease of computation, since everything is done at the null hypothesis.

The outline of the paper is as follows. The general framework and the specific assumptions made are introduced in Data and Likelihoods. The tests are defined explicitly in Score Tests for Familial Correlation. A simulation is given in Simulation. The Example describes the analysis of a subset of the data of Struewing et al. [1997], and suggests that familial correlations may be present. Concluding remarks are given in the Discussion.

DATA AND LIKELIHOODS

Data and Basic Models

In a genotype-proband study, a proband is selected and the phenotype is observed. One then observes the covariates for the proband, as well as the phenotypes and covariates for the other family members. Finally, some of the genotypes of other family members, selected at random, may be observed. In the genotyped-proband design of Struewing et al. [1997], the proband is genotyped. Other designs might also select a random family member, or even the entire family for genotyping.

We make the following assumptions.

- Probands are selected at random within clusters defined by phenotype.
- Alleles are assumed to be in Hardy-Weinberg equilibrium and random mating is assumed.
- The marginal distribution of genotypes does not depend on the covariates, i.e., there is no stratification by genotype.
- The random effects, if any, are independent of the covariates and genotypes.
- The proband's phenotype is conditionally independent of the covariates of the other family members, given the family-level random effect and the proband's genotype and covariates.
- The distribution of genotypes in a family is a known function of the mutant allele frequency q. Gail et al. [1999] describe a simple method of exhaustive enumeration for small pedigrees to compute these mass functions using Mendelian calculations based on the assumptions of Hardy-Weinberg equilibrium and random mating.

Likelihood Functions

The data as described above are a special form of a missing data problem, where some or all of the genotypes in a family are missing, at random, in the nomenclature of Little and Rubin [1987]. By the nature of the genotyped-proband design, since all analysis is done subsequent to the selection of the proband, and the selection probability depends on the phenotype Y^p of the proband, it would ordinarily make sense to compute the likelihood conditioned only on Y^p. As shown in the Appendix, the difficulty with such an analysis is that in order to implement it, one must also specify a joint distribution for all the covariates, and analyses would not be robust to misspecification of this distribution. The analysis would also be made more complex by the addition of more parameters in the joint (marginal) distribution of covariates.

To avoid this difficulty, we have taken an alternative approach, where we condition not only on the phenotype of the proband, but also on the covariates. It is the special feature of this problem, and the assumptions that we have made, that enable us to compute a conditional likelihood that does not depend on the distribution of the covariates. The result is a semiparametric method in the sense of Robins et al. [1994]. While some efficiency may be lost in this way, we doubt that the loss is very great, and it is certainly outweighed by the simplicity and model-robustness of the subsequent analysis.

SCORE TESTS FOR FAMILIAL CORRELATION

Notation and General Definitions

This section gives a precise description of the data and likelihood functions.

The phenotypes of a family are denoted by $\underline{\tilde{\mathbf{Y}}}$, the genotypes by $\underline{\tilde{\mathbf{g}}}$, and covariates by $\underline{\tilde{\mathbf{Z}}}$. The distribution of the phenotype depends on a parameter β, as well as the genotypes and covariates. The random effect denoted by $\zeta = \kappa^{1/2}\nu$, where ν has a distribution function $F_{RE}(\cdot)$ with mean zero and variance one. The proband phenotype is Y^p, and the covariates for the proband are $\underline{\tilde{\mathbf{Z}}}^p$. Finally, some of the genotypes of the family may be observed, which we denote by $\underline{\tilde{\mathbf{g}}}^p$. In the genotyped-proband design of Struewing et al. [1997], $\underline{\tilde{\mathbf{g}}}^p$ is the genotype of the proband. Other designs might also select a random family member, or even the entire family for genotyping. In either case, $\underline{\tilde{\mathbf{g}}}^p$ are the observed genotypes. We will write $\underline{\tilde{\mathbf{g}}}_m$ as the missing genotypes.

The joint distribution for a family of the phenotypes given the covariates, genotype, and random effect is denoted by $f(\underline{\tilde{\mathbf{y}}}|\underline{\tilde{\mathbf{z}}}, \underline{\tilde{\mathbf{g}}}, \zeta, \beta)$.

The conditional mass function of the missing genotypes is denoted by $f_{CG}(\underline{\tilde{\mathbf{g}}}_m|\underline{\tilde{\mathbf{g}}}^p, \underline{\tilde{\mathbf{z}}}, q) = f_{CG}(\underline{\tilde{\mathbf{g}}}_m|\underline{\tilde{\mathbf{g}}}^p, q)$, where "CG" stands for "conditional genotype probability mass function."

The marginal distribution of genotypes does not depend on the covariates, i.e., there is no stratification by genotype. We write this as $f_G(\underline{\tilde{\mathbf{g}}}^p|\underline{\tilde{\mathbf{z}}}, q) = f_G(\underline{\tilde{\mathbf{g}}}^p|q)$, where "G" stands for "genotype probability mass function."

The density or probability mass function of the proband's phenotype given the covariates, observed genotypes, and random effect $f(y^p|\underline{\tilde{\mathbf{z}}}, \underline{\tilde{\mathbf{g}}}^p, \zeta, \beta) = f(y^p|\underline{\tilde{\mathbf{z}}}^p, \underline{\tilde{\mathbf{g}}}^p, \zeta, \beta)$.

Let $\mathcal{B} = (\beta, q)$. A simple analysis shows that the likelihood of the observed data given $(Y^p, \underline{\tilde{\mathbf{Z}}})$ is

$$f(\tilde{\underline{\mathbf{Y}}},\ \tilde{\underline{\mathbf{g}}}^p|Y^p,\tilde{\underline{\mathbf{Z}}},\kappa,\mathcal{B}) = f(\tilde{\underline{\mathbf{Y}}},\tilde{\underline{\mathbf{g}}}^p|\tilde{\underline{\mathbf{Z}}}) \,/\, f(Y^p|\tilde{\underline{\mathbf{Z}}})$$

$$= \frac{f_G(\tilde{\underline{\mathbf{g}}}^p|q)\int\int f(\tilde{\underline{\mathbf{Y}}}|\tilde{\underline{\mathbf{Z}}},\tilde{\underline{\mathbf{g}}},\zeta=\kappa^{1/2}\upsilon,\beta)\, f_{CG}(\tilde{\underline{\mathbf{g}}}_m|\tilde{\underline{\mathbf{g}}}^p,q)\mathrm{d}F_{RE}(\upsilon)\mathrm{d}\mu(\tilde{\underline{\mathbf{g}}}_m)}{\int\int f(Y^p|\tilde{\underline{\mathbf{Z}}}^p,\tilde{\underline{\mathbf{g}}}^p,\zeta=\kappa^{1/2}\upsilon,\beta)\, f_G(\tilde{\underline{\mathbf{g}}}^p|q)\mathrm{d}F_{RE}(\upsilon)\mathrm{d}\mu(\tilde{\underline{\mathbf{g}}}^p)}\ , \tag{1}$$

where the notation $\int \mathrm{d}\mu(\cdot)$ means a summation over the argument.

The special structure of our conditional likelihood enables easy computation of the score test for the hypothesis that $\kappa = 0$. The general theory is fairly standard, and we now outline it. Later we specialize the general theory to our case, showing that considerable simplifications occur. As before, $\mathcal{B} = (\beta, q)$.

General Theory

For our problem, referring to (1), write the score statistic for the ith family as

$$S_\kappa(\tilde{\underline{\mathbf{Y}}}_i,\tilde{\underline{\mathbf{g}}}^p_i,\tilde{\underline{\mathbf{Z}}}_i,\ \mathcal{B}) = (\partial/\partial\kappa)\log\left\{ f(\tilde{\underline{\mathbf{Y}}}_i,\tilde{\underline{\mathbf{g}}}^p_i|Y_i^p,\tilde{\underline{\mathbf{Z}}}_i,\kappa,\ \mathcal{B})\right\}\Big|_{\kappa=0}. \tag{2}$$

Also under the hypothesis, write the score for $\mathcal{B}$ in the ith family as

$$S_\mathcal{B}(\tilde{\underline{\mathbf{Y}}}_i,\tilde{\underline{\mathbf{g}}}^p_i,\tilde{\underline{\mathbf{Z}}}_i,\ \mathcal{B}) = (\partial/\partial\mathcal{B})\log\left\{ f(\tilde{\underline{\mathbf{Y}}}_i,\tilde{\underline{\mathbf{g}}}^p_i|Y_i^p,\tilde{\underline{\mathbf{Z}}}_i,\kappa=0,\ \mathcal{B})\right\}. \tag{3}$$

Under the hypothesis that $\kappa = 0$, the maximum likelihood estimate $\hat{\mathcal{B}}$ will usually be computed numerically. The derivatives (3) can also be computed by numerical differentiation.

For the ith family, define the expected information matrices by

$$\mathcal{I}_{\kappa\kappa i}(\mathcal{B}) = E\left\{S^2_\kappa(\tilde{\underline{\mathbf{Y}}}_i,\tilde{\underline{\mathbf{g}}}^p_i,\tilde{\underline{\mathbf{Z}}}_i,\mathcal{B})|\kappa=0)\right\};$$

$$\mathcal{I}_{\kappa\mathcal{B}i}(\mathcal{B}) = E\left\{S_\kappa(\tilde{\underline{\mathbf{Y}}}_i,\tilde{\underline{\mathbf{g}}}^p_i,\tilde{\underline{\mathbf{Z}}}_i,\mathcal{B})S^t_\mathcal{B}(\tilde{\underline{\mathbf{Y}}}_i,\tilde{\underline{\mathbf{g}}}^p_i,\tilde{\underline{\mathbf{Z}}}_i,\mathcal{B})|\kappa=0\right\};$$

$$\mathcal{I}_{\mathcal{B}\mathcal{B}i}(\mathcal{B}) = E\left\{S_\mathcal{B}(\tilde{\underline{\mathbf{Y}}}_i,\tilde{\underline{\mathbf{g}}}^p_i,\tilde{\underline{\mathbf{Z}}}_i,\mathcal{B})S^t_\mathcal{B}(\tilde{\underline{\mathbf{Y}}}_i,\tilde{\underline{\mathbf{g}}}^p_i,\tilde{\underline{\mathbf{Z}}}_i,\mathcal{B})|\kappa=0\right\},$$

where these expectations are all computed with respect to the density or probability mass function in (1). Note that because we have conditioned properly, these expectations do not involve the *distribution* of the covariates, although they are of course functions of the observed covariates.

Let p be the number of components of $\mathcal{B}$, and let n be the number of families. It follows from standard likelihood theory that the score test statistic is

$$\frac{\{n/(n-p)\}^{1/2}\Sigma_{i=1}^n S_\kappa(\tilde{\underline{\mathbf{Y}}}_i,\tilde{\underline{\mathbf{g}}}^p_i,\tilde{\underline{\mathbf{Z}}}_i,\hat{\mathcal{B}})}{\left[\Sigma_{i=1}^n\mathcal{I}_{\kappa\kappa i}(\hat{\mathcal{B}}) - \{\Sigma_{i=1}^n\mathcal{I}_{\kappa\mathcal{B}i}(\hat{\mathcal{B}})\}\{\Sigma_{i=1}^n\mathcal{I}_{\mathcal{B}\mathcal{B}i}(\hat{\mathcal{B}})\}^{-1}\Sigma_{i=1}^n\mathcal{I}^t_{\kappa\mathcal{B}i}(\hat{\mathcal{B}})\right]^{1/2}}, \tag{4}$$

where the leading term is a correction for degrees of freedom [Simpson et al., 1997]. This is a one-sided test, and thus to achieve a nominal level of α, the test statistic (4) should be compared to the $1 - \alpha$ percentile of the t-distribution with $n - p$ degrees of freedom: of course, if the number of probands n is large, this percentile is essentially the same as the standard normal percentile.

In some problems, the expectations required to compute the terms in the denominator may be too difficult to compute in practice, in which case one could use observed information and remove the expectations in $\mathcal{I}_{kk\ell}(\mathcal{B})$, $\mathcal{I}_{k\mathcal{B}\ell}(\mathcal{B})$ and $\mathcal{I}_{\mathcal{B}\mathcal{B}\ell}(\mathcal{B})$.

Computing the Score Statistic

In order to implement (4), we need to compute the numerator. In this section we derive the formulae for this computation. Using the methods of Liang [1987], and ignoring the indices if all genotypes were known, the score statistic for an entire family or the proband is usually easily computed. We will later give some examples of this computation, but in what follows, if we had data from an entire family (including genotypes), then we will assume that the score statistic can be computed, and we will write it as

$$S_{1,\kappa}(\tilde{\underline{\mathbf{Y}}},\tilde{\underline{\mathbf{g}}},\tilde{\underline{\mathbf{Z}}},\beta) = \{\partial / \partial\kappa\}\log\left\{\int f(\tilde{\underline{\mathbf{Y}}}|\tilde{\underline{\mathbf{Z}}},\tilde{\underline{\mathbf{g}}},\zeta = \kappa^{1/2}\upsilon,\beta)\mathrm{d}F_{RE}(\upsilon)\right\}\Big|_{\kappa=0}. \tag{5}$$

At least formally, the same calculation can be done for data from the proband only, and we will write this as

$$S_{2,\kappa}(Y^p,\tilde{\underline{\mathbf{g}}}^p,\tilde{\underline{\mathbf{Z}}},\beta) = \{\partial / \partial\kappa\}\log\left\{\int f(Y^p|\,\tilde{\underline{\mathbf{Z}}}^p,\tilde{\underline{\mathbf{g}}}^p,\zeta = \kappa^{1/2}\upsilon,\beta)\mathrm{d}F_{RE}(\upsilon)\right\}\Big|_{\kappa=0}.$$

As shown by Liang [1987] and the example considered below, in many important problems $S_{1,\kappa}(\cdot)$ and $S_{2,\kappa}(\cdot)$ have a convenient form.

In the Appendix, we show that knowledge of the score statistic for completely known genotypes leads easily to the following formulae:

$$\begin{aligned}
&S_\kappa(\tilde{\underline{\mathbf{Y}}},\tilde{\underline{\mathbf{g}}}^p,\mathcal{B}) = H_1(\tilde{\underline{\mathbf{Y}}},\tilde{\underline{\mathbf{g}}}^p,\mathcal{B}) - H_2(\tilde{\underline{\mathbf{Y}}},\,\mathcal{B}),\ \text{where}\\
&H_1(\tilde{\underline{\mathbf{Y}}},\tilde{\underline{\mathbf{g}}}^p,\mathcal{B}) = \frac{\int T_1(\tilde{\underline{\mathbf{Y}}},\tilde{\underline{\mathbf{g}}} = (\tilde{\underline{\mathbf{g}}}_m,\tilde{\underline{\mathbf{g}}}^p),\ \tilde{\underline{\mathbf{Z}}},\beta)\, f_{CG}(\tilde{\underline{\mathbf{g}}}_m|\tilde{\underline{\mathbf{g}}}^p,q)\mathrm{d}\mu(\tilde{\underline{\mathbf{g}}}_m)}{\int f(\tilde{\underline{\mathbf{Y}}},\tilde{\underline{\mathbf{g}}} = (\tilde{\underline{\mathbf{g}}}_m,\tilde{\underline{\mathbf{g}}}^p),\ \tilde{\underline{\mathbf{Z}}},\zeta = 0,\beta)\, f_{CG}(\tilde{\underline{\mathbf{g}}}_m|\tilde{\underline{\mathbf{g}}}^p,q)\mathrm{d}\mu(\tilde{\underline{\mathbf{g}}}_m)};\\
&H_2(\tilde{\underline{\mathbf{Y}}},\mathcal{B}) = \frac{\int T_2(Y^p,\tilde{\underline{\mathbf{g}}}^p,\tilde{\underline{\mathbf{Z}}}^p,\beta)\, f_G(\tilde{\underline{\mathbf{g}}}^p|q)\mathrm{d}\mu(\tilde{\underline{\mathbf{g}}}^p)}{\int f(Y^p|\ \tilde{\underline{\mathbf{g}}}^p,\ \tilde{\underline{\mathbf{Z}}}^p,\zeta = 0,\beta)\, f_G(\tilde{\underline{\mathbf{g}}}^p|q)\mathrm{d}\mu(\tilde{\underline{\mathbf{g}}}^p)};\\
&T_1(\tilde{\underline{\mathbf{Y}}},\tilde{\underline{\mathbf{g}}},\tilde{\underline{\mathbf{Z}}},\beta) = S_{1,\kappa}(\tilde{\underline{\mathbf{Y}}},\tilde{\underline{\mathbf{g}}},\tilde{\underline{\mathbf{Z}}},\beta)\, f(\tilde{\underline{\mathbf{Y}}}|\tilde{\underline{\mathbf{g}}},\tilde{\underline{\mathbf{Z}}},\zeta = 0,\beta);\\
&T_2(Y^p,\tilde{\underline{\mathbf{g}}}^p,\tilde{\underline{\mathbf{Z}}}^p,\beta) = S_{2,\kappa}(\tilde{\underline{\mathbf{Y}}}^p,\tilde{\underline{\mathbf{g}}}^p,\tilde{\underline{\mathbf{Z}}}^p,\beta)\, f(Y^p|\tilde{\underline{\mathbf{g}}}^p,\tilde{\underline{\mathbf{Z}}}^p,\zeta = 0,\beta).
\end{aligned} \tag{6}$$

We have, thus, derived an explicit formula for the numerator of (4). It is important to note that, in common with the results of Liang [1987], the score test statistic does not require that we actually specify the non-null distribution of the random effects.

Binary and Weibull Phenotypes, Classical Genotyped-Proband Design

Consider an autosomal dominant gene in the genotyped-proband design with a binary phenotype and no covariates. The natural generalization to allow for random effects is to set pr$\{Y = 1|G \in (1,2), \zeta\} = H(\gamma + \zeta)$ and pr$(Y = 1|G = 3, \zeta) = H(\eta + \zeta)$, where as before $\zeta = \kappa^{1/2}\nu$, where ν has a distribution function $F_{RE}(\cdot)$ with mean zero and variance one. We want to test that $\kappa = 0$, i.e., that there is no familial correlation other than that determined by the family's genotypes. For this problem, all the terms necessary to form the score test are easily computed, and are available from the first author.

For survival times, we use the improper Weibull model of Gail et al. [1999], in which the penetrance is expressed as the probability that an individual will eventually become diseased. In the absence of a random effect, for a person with genotype g, we assume that the survival function is

$$S_g(t) = 1 - \phi_g + \phi_g \exp\ (-\lambda_g^{\alpha_g} \tau^{\alpha_g}). \tag{7}$$

The penetrances ϕ_0 and ϕ_1 correspond to lifetime risks for those not carrying and carrying the gene, respectively. The improper survival distribution (7) arises as a mixture of ϕ_g susceptible and $1 - \phi_g$ non-susceptible members of the population, the former having a Weibull distribution of times to disease.

The hazard function for susceptibles is $\lambda_g^{\alpha_g}\alpha_g t^{\alpha g-1}$, and it is by multiplying this hazard by a "frailty" factor that we will incorporate a random effect. In fact, given a familial random effect $\zeta = \kappa^{1/2}\nu$, the survival functions and density functions for susceptibles given the random effects are

$$S_g(t) = 1 - \phi_g + \phi_g \exp\ \{-\lambda_g^{\alpha_g} t^{\alpha_g}\ \exp(\kappa^{1/2}\nu)\};$$

$$f_g(t) = \alpha_g \phi_g \lambda_g^{\alpha_g} t^{\alpha_g - 1}\ \exp(\kappa^{1/2}\nu)\ \exp\{-\lambda_g^{\alpha_g} t^{\alpha_g}\ \exp(\kappa^{1/2}\nu)\}.$$

For this problem, all the terms necessary to form the score test are computed in the Appendix.

SIMULATION

We performed a small simulation to illustrate the results with a binary phenotype and an autosomal dominant gene. Among carriers of the gene, the probability of disease was $\phi_1 = 0.92$, this being the penetrance of the gene. Among non-carriers, the probability of disease was set to $\phi_0 = 0.10$. The allele rate was set to $q = 0.0033$. As described by Gail et al. [1999], the values of (q, ϕ_0, ϕ_1) were chosen to reflect penetrances and allele frequencies estimated by Claus et al. [1991], who studied the risk from a hypothetical autosomal dominant mutation for breast cancer.

The power of the score test was computed for a random effects model with $\zeta = \kappa^{1/2}\nu$, where ν is a standard normal random variable. We assumed a family of size 3, consisting of a mother, a sister, and a sister-proband. For each family, genotypes were generated under our assumptions, and then a family-level random effect ζ was generated. Then among carriers of the gene, phenotypes followed a Bernoulli model with probability of disease $H(\gamma + \kappa^{1/2}\nu)$, where $H(\cdot)$ is the logistic distribution function and γ was chosen so that pr(disease | carrier) = $E(H(\gamma + \kappa^{1/2}\nu) = \phi_1$. Similarly, among non-carriers of the gene, phenotypes followed a Bernoulli model with probability of disease $H(\eta + \kappa^{1/2}\nu)$, where $H(\cdot)$ is the logistic distribution function and η was chosen so that pr(disease | non-carrier) = $EH(\eta + \kappa^{1/2}\nu) = \phi_0$. We generated a large number of such families, and then selected 400 families with a case-proband, and 400 families with a control proband.

We simulated the score test 500 times for the values $\kappa^{1/2}$ = 0.0 (null case), 0.3, 0.6, 1.0. In the null case, the parameter estimates were found to be nearly unbiased (see Table I), but the maximum likelihood estimate of the penetrance ϕ_1 equaled 1.0 approximately 35% of the time. The simulated level of a nominal 5%-level test was found to be 0.048, very near the nominal level.

The power of the test for $\kappa^{1/2}$ = 0.3, 0.6, 1.0 was found to be 0.140, 0.724, 1.000, respectively. To understand the magnitude of these alternatives, we constructed what we call the "family-influence," namely

$$\text{family - influence} = \frac{\text{pr}(Y_{sister} = 1 | Y_{proband} = 1, g_{proband} = 1, \zeta = \kappa^{1/2}\upsilon)}{\text{pr}(Y_{sister} = 1 | Y_{proband} = 1, g_{proband} = 1)}.$$

This influence is displayed as a function of v in Figure 1. We see here, for example, that when κ = 1.0, the chance that a sister has the disease given that the proband is a carrier and has the disease is increased by approximately 50% for a very large random effect (2.5 standard deviation). This increased influence is only about 15% when the random effect is 1.0 standard deviation. We believe that the results indicate that the score test can detect a large departure from the conditional independence assumption.

One might also test for familial aggregation not related to the gene under study

TABLE I. Results From a Simulation of a Binary Phenotype for Possible Family Level Correlations in Phenotypes When the Allele Frequency is q = 0.0033, the Penetrance for Wild Type is θ_0 = 0.10 and the Penetrance for Mutant Type is θ_1 = 0.92

Random effect s.d.	0.0	0.3	0.6	1.0	2.0	4.0
Mean (q)	0.0034	0.0055	0.0058	0.0044	0.0072	0.0132
100× s.e. of mean	0.0036	0.0035	0.0038	0.0045	0.0067	0.0103
Median (q)	0.0033	0.0034	0.0037	0.0044	0.0071	0.0120
Mean (θ_0)	0.100	0.103	0.113	0.133	0.204	0.269
100× s.e. of mean	0.035	0.035	0.036	0.039	0.051	0.051
Median (θ_0)	0.100	0.103	0.112	0.133	0.203	0.268
Mean (θ_1)	0.906	0.910	0.916	0.934	0.962	0.951
s.e. of mean	0.0042	0.0044	0.0042	0.0037	0.0028	0.0018
Median (θ_1)	0.922	0.941	0.948	0.980	1.000	1.000

*There were 500 simulated experiments, each with 400 proband cases and 400 proband controls.

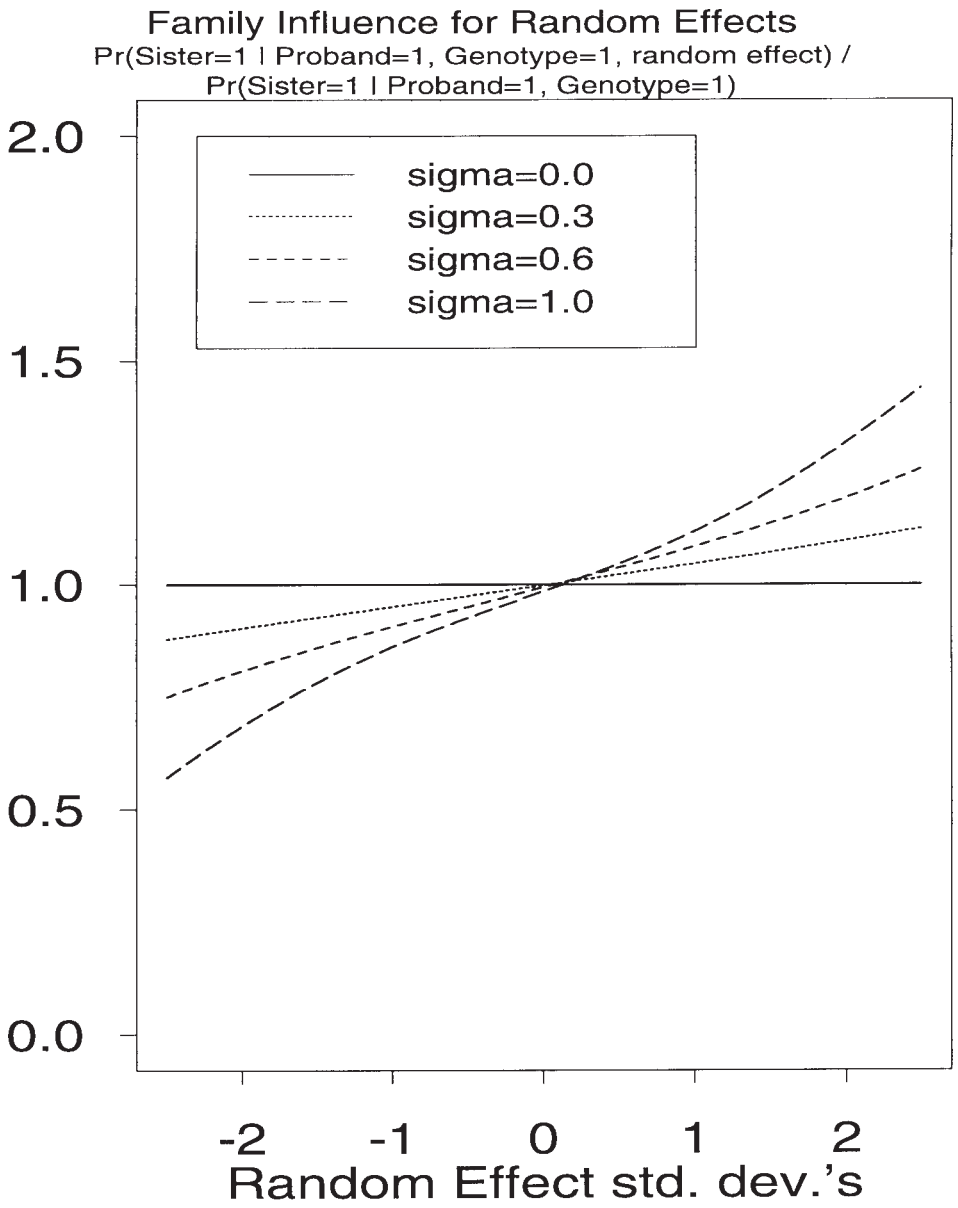

Fig. 1. Let the random effect be $\zeta = \kappa^{1/2}v$. The family consists of a mother, a sister and a sister-proband. This is $\text{pr}(Y_{sister} = 1|Y_{proband} = 1, g_{proband} = 1, \zeta)/\text{pr}(Y_{sister} = 1|Y_{proband} = 1, g_{proband} = 1)$ as a function of $x = v = \zeta/\kappa^{1/2}$, i.e., as a function of standard deviations of the random effect.

by examining correlations among phenotypes of first-degree relatives of non-carrier probands. With rare allele frequencies, the gene under study should contribute very little to such correlations. We tested for intraclass correlation between the sister and the mother of non-carrier probands. We first subtracted the sister and mother means from the responses and then did a standard permutation test of the product. In our

simulations, these tests had a level slightly higher than the nominal 5% (level = 0.072), and their power was lower than that of the score test (power = 0.124, 0.240, 0.734, respectively).

The fact that the score test has more power than ad hoc tests such as described above is to be expected, since the score test is known to be the (locally) most powerful test for familial correlation.

Effects of Residual Family Correlation

As part of the simulations, we were able to examine the effect of residual family correlation on the parameter estimates of allele frequency q, penetrance for mutant type ϕ_0, and penetrance for rare type ϕ_1. The results are given in Table I when the family random effects have standard deviation $\kappa^{1/2}$ = 0.0, 0.3, 0.6, 1.0, 2.0, 4.0. The most noticeable result is that for increasingly more severe random effects, all parameters are overestimated. Even when the standard deviation of this random effect equals 1.0, the allele frequency is overestimated by 1/3, and it is overestimated by 100% when the standard deviation is as large as 2.0.

The over-estimation of all three parameters appears to be a small-sample phenomenon. For other purposes we repeated the simulations with $\kappa^{1/2}$ = 1.0 but with 9.750 proband controls and 1,083 proband cases, and found that while q and ϕ_1 were still overestimated with simulation means 0.004 and 0.938, respectively, ϕ_0 was slightly underestimated with simulation mean 0.098.

EXAMPLE

We were able to obtain a subset of the data considered by Struewing et al. [1997]. First, all families were identified that had known breast cancer status for a proband, at least one sister and her mother. Then, a randomly selected sister of the proband was chosen. The final data set made available to us then consisted of 1,960 families with a proband, her mother and a single sister. We based our analysis upon the Weibull model (7). There were 143 case probands.

Under the conditional independence assumption, the lifetime penetrance for those having the gene was estimated as ϕ_1 = 0.68, while for those without the gene the penetrance was estimated as ϕ_0 = 0.27. The allele frequency was estimated as q = 0.0122. The score test statistic had a value of 2.36, which has a one-sided significance level of 0.009. Thus, the evidence points to a violation of the conditional independence assumption. To check that the level of the test was nearly the nominal, and thus that the effect we are observing is not due to problems with the test statistic, we ran a small parametric bootstrap simulation. Data were generated according to the model (7), with no random effects and parameters set to their maximum likelihood estimates. We ran 100 simulations to test this null model, and at the nominal 5% level, observed 5 rejections.

Several possibilities could account for residual "non-genetic" correlation in these data. Other mutations than the BRCA1 and BRCA2 mutations under study could be segregating in these families, such as genetic factors that influence reproductive hormone levels. Shared diet or other environmental factors could also account for residual correlation. Selection bias is another possibility. In particular, if a subject is more likely to volunteer to be a proband in the study if her mother and sister both have breast cancer, an artifactual correlation could be induced.

DISCUSSION

While we have focused on the score test for a random familial effect in genotyped-proband studies, the methods derived are actually an example of a more general phenomenon. In the Appendix, we consider the case that one of the covariates is sometimes missing, while the other covariates are always observable. This is a form on monotone missingness [Little and Rubin, 1997]. In the problem we have considered, the partially missing covariate is genotype, while all other covariates were observable. Our analysis conditioned on these always observable covariates. Conditioning on the covariates that are never missing, it is generally easy to compute the score test for a random effect if the conditional distribution of the missing covariates can be specified as a function of the observed covariates and a parameter. This result could be useful if, in addition to genotypes, other covariates are missing, such as age at menarche or age at first live birth. As we show in the Appendix, in order to compute the score test, one must specify the distribution of the missing covariates given the observed ones as in a standard likelihood analysis. This may be more or less easy and practicable in different situations.

More generally, if missing data were not monotone, e.g., those covariates other than the genotype that we have assumed are always observable are instead sometimes missing, a full likelihood analysis can be contemplated. However, as indicated in the Appendix, this requires the specification of the joint distribution of the covariates within a family. This is possible to do, but care must be taken to specify an appropriate distribution that allows for correlations of the covariates within a family.

Our discussion, simulation, and example have focused on the use of phenotypes of *first*-degree relatives. However, the formulae are sufficiently general to allow for observing the phenotypes of second-degree relatives, as long as the random effect is now interpreted as belonging to an extended family.

ACKNOWLEDGMENTS

R.J. Carroll's research was supported by a grant from the National Cancer Institute (CA-57030), and by the Texas A&M Center for Environmental and Rural Health via a grant from the National Institute of Environmental Health Sciences (P30-ESO9106).

APPENDIX

Verification of (6)

Here we verify (6). Referring to (2), it suffices to consider the denominator of (1), since the numerator follows in a similar manner. We have that

$$(\partial / \partial\kappa) \log \left\{ \int\int f(Y^p \left| \underline{\tilde{\mathbf{Z}}}^p, \underline{\tilde{\mathbf{g}}}^p, \zeta = \kappa^{1/2}\upsilon, \beta) f_G(\underline{\tilde{\mathbf{g}}}^p \right| q) \mathrm{d}F_{RE}(\upsilon) \mathrm{d}\mu(\underline{\tilde{\mathbf{g}}}^p) \right\} \Big|_{\kappa=0}$$

$$= \frac{(\partial / \partial\kappa) \int\int f(Y^p \left| \underline{\tilde{\mathbf{Z}}}^p, \underline{\tilde{\mathbf{g}}}^p, \zeta = \kappa^{1/2}\upsilon, \beta) f_G(\underline{\tilde{\mathbf{g}}}^p | q) \mathrm{d}F_{RE}(\upsilon) d\mu(\underline{\tilde{\mathbf{g}}}^p)}{\int\int f(Y^p \left| \underline{\tilde{\mathbf{Z}}}^p, \underline{\tilde{\mathbf{g}}}^p, \zeta = \kappa^{1/2}\upsilon, \beta) f_G(\underline{\tilde{\mathbf{g}}}^p | q) \mathrm{d}F_{RE}(\upsilon) d\mu(\underline{\tilde{\mathbf{g}}}^p)} \Big|_{\kappa=0}$$

$$= \frac{\int \left\{ (\partial / \partial\kappa) \int f(Y^p \left| \underline{\tilde{\mathbf{Z}}}^p, \underline{\tilde{\mathbf{g}}}^p, \zeta = \kappa^{1/2}\upsilon, \beta) \mathrm{d}F_{RE}(\upsilon) \right\} \Big|_{\kappa=0} f_G(\underline{\tilde{\mathbf{g}}}^p | q) \mathrm{d}\mu(\underline{\tilde{\mathbf{g}}}^p)}{\int\int f(Y^p \left| \underline{\tilde{\mathbf{Z}}}^p, \underline{\tilde{\mathbf{g}}}^p, \zeta = 0, \beta) f_G(\underline{\tilde{\mathbf{g}}}^p | q) \mathrm{d}\mu(\underline{\tilde{\mathbf{g}}}^p)}.$$

This last term is easily shown to equal $H_2\underline{\tilde{\mathbf{Y}}}, \underline{\tilde{\mathbf{g}}}^p, \mathcal{B})$ as claimed.

Formulae for the Weibull Model

Consider an individual family with J members. We write $\Delta_j = 1$ if the jth family member is uncensored, and $\Delta_j = 0$ otherwise. Let T_j be the minimum of the survival and censoring times, and let $M = \Sigma_{j=1}^{J} \Delta_j$ be the number of uncensored family members. Define

$$a_0 = \prod_{j=1}^{J} \left\{ \phi_{g_j} \lambda_{g_j}^{\alpha_{g_j}} \alpha_{g_j} T_j^{\alpha_{g_j}-1} \right\}^{\Delta_j};$$

$$a_1 = \prod_{j=1}^{J} \Delta_j \lambda_{g_j}^{\alpha_{g_j}} T_j^{\alpha_{g_j}}.$$

Given the random effect, the likelihood of the uncensored family members is thus

$$\alpha_0 \exp(M\kappa^{1/2}\nu)\exp\{-\alpha_1 \exp(\kappa^{1/2}\nu)\}.$$

Now consider the censored family members. Given the random effects, the likelihood of these censored observations is

$$\prod_{j=1}^{J} \left[1 - \phi_{g_j} + \phi_{g_j} \exp\{-\lambda_{g_j}^{\alpha_{g_j}} T_j^{\alpha_{g_j}} \exp(\kappa^{1/2}\nu)\} \right]^{1-\Delta_j}$$

$$= \sum_{l_1=0}^{1} \cdots \sum_{l_J=0}^{1} \prod_{j=1}^{J} \left[(1-\phi_{g_j})^{l_j} \phi_{g_j}^{1-l_j} \exp\{-(1-l_j)\lambda_{g_j}^{\alpha_{g_j}} T_j^{\alpha} \exp(\kappa^{1/2}\nu)\} \right]^{1-\Delta_j}$$

$$= \sum_{l_1=0}^{1} \cdots \sum_{l_J=0}^{1} b(l_1,\ldots,l_J) \exp\{-c(l_1,\ldots,l_J)\exp(\kappa^{1/2}\nu)\},$$

where

$$b(l_1,\ldots,l_J) = \prod_{j=1}^{J} \left\{ (1-\phi_{g_j})^{l_j} \phi_{g_j}^{1-l_j} \right\}^{1-\Delta_j};$$

$$c(l_1,\ldots,l_J) = \sum_{j=1}^{J} \left\{ (1-l_j)\lambda_{g_j}^{\alpha_{g_j}} T_j^{\alpha} \right\}^{1-\Delta_j}.$$

In what follows, we will write $c(\cdot)$ as a shorthand for $c(\ell_1,...,\ell_J)$, and similarly for $b(\cdot)$. If all the genotypes were observable, then, as we now show, the contribution of this family to the score test statistic, namely (5), is

$$S_{1,\kappa}(\underline{\tilde{\mathbf{Y}}}, \underline{\tilde{\mathbf{g}}}, \underline{\tilde{\mathbf{Z}}}, \beta) = \frac{1}{2} \frac{\Sigma_{l_1=0}^{1} \cdots \Sigma_{l_J=0}^{1} \left[\{c(\cdot) + a_1 - M\}^2 - \{c(\cdot) + a_1\} \right] b(\cdot) \exp\{-c(\cdot)\}}{\Sigma_{l_1=0}^{1} \cdots \Sigma_{l_J=0}^{1} b(\cdot) \exp\{-c(\cdot)\}}. \quad (8)$$

The contribution (6) for an individual is calculated similarly, except that $J = 1$ and calculations are done only for the proband: note that all the terms making up (8) have to be redefined, i.e., M, J, α_1, etc.

To see (8), note that the likelihood given the random effects is

$$H(v,\kappa) = a_0 \exp(M\kappa^{1/2}v)\sum_{\ell_1=0}^{1}\cdots\sum_{\ell_J=0}^{1} b(\ell_1,\ldots,\ell_J)\exp\left[-\{c(\ell_1,\ldots,\ell_J)+a_1\}\exp(\kappa^{1/2}v)\right].$$

The score statistics for this family is

$$\frac{\partial}{\partial\kappa}\log\left\{\int H(\upsilon,\kappa)F_{RE}(\upsilon)\right\}\Big|_{\kappa=0}.$$

This is easily seen to be

$$\lim_{\kappa\to 0}\frac{\int v\exp(M\kappa^{1/2}v)\sum_{\ell_1=0}^{1}\cdots\sum_{\ell_J=0}^{1} d(\ell_1,\ldots\ell_J,\kappa,\upsilon)F_{RE}(\upsilon)}{2\kappa^{1/2}\sum_{\ell_1=0}^{1}\cdots\sum_{\ell_J=0}^{1} b(\cdot)\exp\left[-\{c(\cdot)+a_1\}\right]},$$

where

$$d(\ell_1,\ldots\ell_J,\kappa,\upsilon) = b(\cdot)\exp\left[-\{c(\cdot)+a_1\}\exp(\kappa^{1/2}\upsilon)\right]\left[M-\{c(\cdot)+a_1\}\exp(\kappa^{1/2}\upsilon)\right].$$

Applying L'Hospital's rule, this is seen to equal (8).

Why One Should Condition on Covariates

The likelihood (1) conditions on the covariates $\underset{\sim}{\tilde{\mathbf{Z}}}$ in the family. The advantage of this conditioning is that while the values of the covariates in the sample are important, their distribution is not required. We now argue that for reasons of model robustness, the conditional likelihood (1) is often preferable. We make our argument under the assumption of no family-level random effect, so that $\zeta = 0$.

The genotyped-proband design is based on the proband phenotype Y^p, and hence the likelihood of all the data is that of $(\underset{\sim}{\tilde{\mathbf{Y}}}, \underset{\sim}{\tilde{\mathbf{Z}}}, \underset{\sim}{\tilde{\mathbf{g}}}^p)$ given Y^p. Let $f_Z(\underset{\sim}{\tilde{\mathbf{z}}})$ and $fz^*(\underset{\sim}{\tilde{\mathbf{z}}}^p)$ be the density/mass function of $\underset{\sim}{\tilde{\mathbf{Y}}}$ and $\underset{\sim}{\tilde{\mathbf{Z}}}^p$, respectively. Then the likelihood of all the data is (1) times the term

$$f(\underset{\sim}{\tilde{\mathbf{Z}}}\,|\,Y^p) = f(Y^p\,|\,\underset{\sim}{\tilde{\mathbf{Z}}})\,f_Z(\underset{\sim}{\tilde{\mathbf{Z}}}) / f(Y^p)$$

$$= \frac{f_Z(\underset{\sim}{\tilde{\mathbf{Z}}})\int f(Y^p\,|\,\underset{\sim}{\tilde{\mathbf{Z}}}^p, \underset{\sim}{\tilde{\mathbf{g}}}^p, \zeta=0,\beta)\,f_G(\underset{\sim}{\tilde{\mathbf{g}}}^p\,|\,q)\,\mathrm{d}\mu(\underset{\sim}{\tilde{\mathbf{g}}}^p)}{f(Y^p\,|\,\underset{\sim}{\tilde{\mathbf{z}}}^p, \underset{\sim}{\tilde{\mathbf{g}}}^p, \zeta=0,\beta)\,f_G(\underset{\sim}{\tilde{\mathbf{g}}}^p\,|\,q)\,f_{\underset{\sim}{\tilde{\mathbf{z}}}}(d\mu(\underset{\sim}{\tilde{\mathbf{g}}}^p)d\mu(\underset{\sim}{\tilde{\mathbf{z}}})} \qquad (9)$$

As seen in the denominator of (9), the extra contribution to the likelihood, which comes from *not* conditioning on the covariates $\underset{\sim}{\tilde{\mathbf{Z}}}$, requires that one specify a model for the marginal distribution of the covariates.

Score Tests With Missing Data

Suppose that $\tilde{\underline{\mathbf{Y}}}$ is a response and $\tilde{\underline{\mathbf{W}}}$ a set of covariates, some of which are observed ($\tilde{\underline{\mathbf{W}}}_O$) and some of which are missing ($\tilde{\underline{\mathbf{W}}}_M$). Suppose that a joint density of $\tilde{\underline{\mathbf{W}}}$ depends on a parameter q, and write the density/mass function of the missing covariates given the observed ones as $f_{Z_m|Z_O}(\tilde{\underline{\mathbf{w}}}_M|\tilde{\underline{\mathbf{w}}}_O, q)$. Write the random effects as $\zeta = \kappa^{1/2}\nu$, and write the model density depending on a parameter β as $f(\tilde{\underline{\mathbf{y}}}|\tilde{\underline{\mathbf{w}}}_O, \tilde{\underline{\mathbf{w}}}_M, \zeta, \beta)$. Write $\mathcal{B} = (\beta, q)$. Using the work of Liang [1987], it is often easy to compute the complete-data score statistic.

$$S_\kappa(\tilde{\underline{\mathbf{Y}}}, \tilde{\underline{\mathbf{W}}}_O, \tilde{\underline{\mathbf{W}}}_M, \beta) = (\partial/\partial\kappa)\log\left\{\int f(\tilde{\underline{\mathbf{Y}}}|\tilde{\underline{\mathbf{W}}}_O, \tilde{\underline{\mathbf{W}}}_M, \zeta = \kappa^{1/2}\upsilon, \beta)\mathrm{d}F_{RE}(\upsilon)\right\}\Big|_{\kappa=0}.$$

Then the score statistic for the observed data is easily seen to be

$$(\partial/\partial\kappa)\log\left\{\int\int \mathrm{f}(\tilde{\underline{\mathbf{Y}}}|\tilde{\underline{\mathbf{W}}}_\mathrm{O}, \tilde{\underline{\mathbf{W}}}_\mathrm{M}, \zeta = \kappa^{1/2}\upsilon, \beta)\mathrm{dF}_{\mathrm{RE}}(\upsilon)\mathrm{f}_{\mathrm{Zm|Zo}}(\tilde{\underline{\mathbf{w}}}_\mathrm{M}|\tilde{\underline{\mathbf{W}}}_\mathrm{O}, \mathrm{q})\mathrm{d}\mu(\tilde{\underline{\mathbf{w}}}_\mathrm{M})\right\}\Big|_{\kappa=0}$$

$$= \frac{\int \mathrm{S}_\kappa(\tilde{\underline{\mathbf{Y}}}, \tilde{\underline{\mathbf{W}}}_\mathrm{O}, \tilde{\underline{\mathbf{W}}}_\mathrm{M}, \beta)\mathrm{f}(\tilde{\underline{\mathbf{Y}}}|\tilde{\underline{\mathbf{W}}}_\mathrm{O}, \tilde{\underline{\mathbf{W}}}_\mathrm{M}, \zeta = 0, \beta)\mathrm{f}_{\mathrm{Zm|Zo}}(\tilde{\underline{\mathbf{w}}}_\mathrm{M}|\tilde{\underline{\mathbf{W}}}_\mathrm{O}, \mathrm{q})\mathrm{d}\mu(\tilde{\underline{\mathbf{w}}}_\mathrm{M})}{\int \mathrm{f}(\tilde{\underline{\mathbf{Y}}}|\tilde{\underline{\mathbf{W}}}_\mathrm{O}, \tilde{\underline{\mathbf{W}}}_\mathrm{M}, \zeta = 0, \beta)\mathrm{f}_{\mathrm{Z\,m|Z_O}}(\tilde{\underline{\mathbf{w}}}_\mathrm{M}, \tilde{\underline{\mathbf{W}}}_\mathrm{O}, \mathrm{q})\mathrm{d}\mu(\tilde{\underline{\mathbf{w}}}_\mathrm{M})}. \quad (10)$$

In our case, $\tilde{\underline{\mathbf{W}}}$ consists of the covariates $\tilde{\underline{\mathbf{Z}}}$ and the genotypes $\tilde{\underline{\mathbf{g}}}$. The missing covariates are the genotypes of family members, so that the joint distribution of $\tilde{\underline{\mathbf{Z}}}$ and $\tilde{\underline{\mathbf{W}}}_M$ is completely specified by the allele frequency q, and the integrals in (10) are, thus, easy to compute by summation. Equation (6) is simply then a simple application of the basic idea in (10).

REFERENCES

Claus EB, Risch NJ, Thompson WD. 1991. Genetic analysis of breast cancer in the Cancer and Steroid Hormone Study. Am J Hum Genet 48:232–42.

Commenges D, Jacqmin-Gadda H. 1997. Generalized score test of homogeneity based on correlated random effects models. J R Stat Soc Series B 59:157–70.

Commenges D, Letenneur L, Jacqmin H, Moreau T, Dartigues J. 1994. Test of homogeneity of binary data with explanatory variables. Biometrics 50:613–20.

Commenges D, Jacqmin H, Letenneur L, Van Duijn CM. 1995. Score test for familial aggregation in probands studies: application to Alzheimer's disease. Biometrics 51:542–51.

Gail MH, Benichou J, Pee D, Wacholder S, Carroll RJ. 1999. Genotyped proband design for estimating phenotypic effects. Genet Epidemiol 16:15–39.

Li H, Thompson E. 1997. Semiparametric estimation of major gene and family-specific random effects for age of onset. Biometrics 53:282–93.

Liang KY. 1987. A locally most powerful test for homogeneity with many strata. Biometrika 74:259–64.

Lin X. 1997. Variance component testing in generalized linear models with random effects. Biometrika 84:309–26.

Little RJA, Rubin DB. 1987. The analysis of missing data. New York: John Wiley and Sons.

Robins JM, Rotnitzky A, Zhao LP. 1994. Estimation of regression coefficients when some regressors are not always observed. J Am Stat Assoc 89:846–66.

Simpson DG, Guth D, Zhou H, Carroll RJ. 1997. Interval censoring and marginal analysis in ordinal regression. J Agric Biol Environ Stat 1:354–76.

Struewing JP, Hartge P, Wacholder S, Baker SM, Berlin M, McAdams M, Timmeman MM, Brody LC, Tucker MA. 1997. The risk of cancer associated with specific mutations of BRCA1 and BRCA2 among Ashkenazi Jews. New Engl J Med 336:1401–8.

Wacholder S, Hartge P, Struewing JP, Pee D, Brody L, Tucker MA. 1998. The kin-cohort study for estimating penetrance. Am J Epidemiol 148:623–30.